



# Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire - Constructions impersonnelles et expressions verbales figées

Benoît Sagot, Laurence Danlos

## ► To cite this version:

Benoît Sagot, Laurence Danlos. Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire - Constructions impersonnelles et expressions verbales figées. Cahiers du Cental, 2008, Description linguistique pour le traitement automatique du français, 5, pp.107-126. inria-00515460

**HAL Id: inria-00515460**

**<https://inria.hal.science/inria-00515460>**

Submitted on 7 Sep 2010

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire

## *Constructions impersonnelles et expressions verbales figées*

Benoît Sagot<sup>1</sup>, Laurence Danlos<sup>2</sup>  
INRIA Futurs (projet Signes), Université Paris 7 (Lattice)

### Abstract

We intend to develop a large-coverage morphological and syntactic lexicon for French which can be directly used in Natural Language Processing (NLP) applications, in particular in those involving deep parsing, regardless of the underlying grammatical framework. This lexicon, named *Lefff* (Lexique des Formes Fléchies du Français — Lexicon of French inflected Forms), has been under development since 2004. At the beginning, this lexicon contained only verbal morphological information, mostly automatically induced from corpora. It now covers all parts of speech, and is progressively enriched with syntactic information. In this paper, we show how we used the lexicon-grammar tables, whose development has been initiated by M. Gross, to enrich the *Lefff*. These tables are a valuable starting point. However, it is necessary to achieve both a linguistic and formal modeling work, in order to exploit their content in a NLP lexicon such as the *Lefff*. We illustrate this approach on two kinds of non-standard verbal and adjectival entries : impersonal structures and verbal idiomatic expressions.

**Keywords** : Lexicon-grammar, *Lefff*, impersonal constructions, verbal idiomatic expressions

### Résumé

Nous avons le projet de développer un lexique morphologique et syntaxique du français à large couverture qui soit directement utilisable dans des applications de Traitement Automatique du Langage (TAL), en particulier celles nécessitant une analyse syntaxique profonde, et ce, quel que soit le cadre théorique utilisé. Ce lexique, baptisé *Lefff* (Lexique des Formes Fléchies du Français), est en cours de développement depuis 2004. Au départ, ce lexique ne comportait que des informations morphologiques verbales, principalement induites automatiquement à partir de corpus. Il couvre maintenant toutes les catégories, et est progressivement enrichi d'informations syntaxiques. Nous montrons ici comment nous l'avons enrichi à partir des tables du lexique-grammaire, initialement développées autour de M. Gross. Celles-ci constituent un point de départ d'une valeur inestimable. Il est néanmoins nécessaire de procéder à un double travail de linguistique et de modélisation, afin d'exploiter leur contenu dans un lexique TAL tel que le *Lefff*. Nous illustrons cette approche sur deux types d'entrées verbales et adjectivales non standard : les constructions impersonnelles et les expressions verbales figées.

**Mots-clés** : Lexique-grammaire, *Lefff*, constructions impersonnelles, expressions verbales figées

---

<sup>1</sup>INRIA Futurs, projet Signes, benoit.sagot@inria.fr

<sup>2</sup>Université Paris 7, Institut Universitaire de France, Lattice, laurence.danlos@linguist.jussieu.fr

## 1. Introduction

L'analyse syntaxique profonde ne peut se faire qu'à la condition de disposer d'informations lexicales riches qui forment un lexique morphologique et syntaxique du français à large couverture. De plus, un tel lexique doit représenter ces informations d'une façon adaptée à l'utilisation dans les systèmes de traitement automatique.

Une telle ressource n'est pas disponible pour le français, malgré de nombreuses initiatives indépendantes, qui ont atteint des degrés d'avancement divers et qui ont privilégié des aspects différents. Deux d'entre elles ont donné naissance à des ressources à large couverture que tout sépare :

- Les tables du lexique-grammaire, dont le développement a été initié par Maurice Gross au LADL (Gross 1975) et se poursuit à l'IGM autour d'Éric Laporte,
- Le *Lefff* (Lexique des Formes Fléchies du Français), lexique morphologique et syntaxique du français à large couverture (520 000 entrées) utilisé dans divers systèmes de TAL (Sagot *et al.* 2006).

Il existe une troisième ressource lexicale pour le français, qui ne couvre que les verbes simples, mais dont les fondements linguistiques sont très solides. Il s'agit du lexique DICOVALENCE, successeur du dictionnaire PROTON, développé par Karel van den Eynde et Piet Mertens (van den Eynde et Mertens 2006). Cette ressource, développée dans le cadre de l'approche pronominale initiée par Claire Blanche-Benvéniste et Karel van den Eynde (van den Eynde et Blanche-Benvéniste 1978), est sur le point d'être mise à la disposition de la communauté. Nous en reparlerons dans la suite de cet article.

L'intérêt principal des tables du lexique-grammaire réside dans leur qualité: elles sont le résultat de travaux minutieux et précis d'investigation linguistique menés depuis les années 1970. Par conséquent, l'ensemble des informations qu'elles représentent constitue une source d'informations riche et précise. Cependant, ces informations ne sont pas directement exploitables dans des systèmes d'analyse, une partie importante de ces informations n'étant compréhensible que moyennant de nombreuses connaissances implicites. À l'inverse, l'intérêt principal du *Lefff* est que sa structure et son format sont spécifiquement adaptés à une utilisation dans des systèmes de TAL. Mais la qualité, la richesse et la couverture du *Lefff* n'atteignent pas, loin s'en faut, celles des tables du lexique-grammaire.

Nous avons donc étudié le moyen de profiter simultanément de la couverture et de la précision des tables du lexique-grammaire et de l'adéquation du *Lefff* aux traitements automatiques. Parallèlement à d'autres travaux complémentaires, qui cherchent à comparer le *Lefff* à un lexique au même format généré à partir des tables du lexique-grammaire (Sagot et Gardent, *en cours*), nous avons travaillé à l'utilisation directe des tables du lexique-grammaire pour valider, invalider, ou compléter certaines informations présentes dans le *Lefff*.

Nous avons insisté en particulier sur un certain nombre d'entrées, de constructions et d'informations mal représentées dans les lexiques syntaxiques en général et le *Lefff* en particulier, à savoir :

- les entrées pour les constructions impersonnelles,
- les expressions verbales figées.

Ces travaux sont une étape supplémentaire pour faire à terme du *Lefff* un lexique syntaxique de référence pour le français à l'aide de cette ressource exceptionnelle que sont les tables du lexique-grammaire.

Après une description des tables du lexique-grammaire, et en particulier des tables concernant les constructions impersonnelles et les expressions verbales figées, nous présenterons le *Lefff* plus en détails. La dernière section est consacrée à la modélisation dans le *Lefff* des données linguistiques sur ces deux types de constructions verbales, telles qu'elles peuvent être extraites des tables du lexique-grammaire.

## 2. Le lexique-grammaire : constructions impersonnelles et expressions verbales figées

### 2.1. Brève introduction au lexique-grammaire des verbes standard

Le lexique-grammaire est composé d'un ensemble de tables : 61 tables pour les verbes (la catégorie la mieux décrite), environ 30 tables pour les adjectifs (travail en cours d'achèvement) et plusieurs tables pour les noms dits « prédicatifs » (noms avec argument(s) qui sont étudiés avec leur verbe support). Chaque table regroupe les éléments d'une catégorie donnée partageant une « propriété définitoire ». Une table se présente sous forme de matrice : en lignes, les éléments lexicaux de la table ; en colonnes, les propriétés qui ne sont pas forcément respectées par tous les éléments de la table ; à la croisée d'une ligne et d'une colonne le signe + ou - en suivant une sémantique évidente. Nous allons commencer par présenter les propriétés définitoires des classes avant d'aborder leur aspect matriciel.

Les propriétés définitoires relèvent généralement du cadre de sous-catégorisation. Ainsi, les critères les plus communément utilisées dans les propriétés définitoires sont :

- le nombre de compléments (la notion de complément utilisée ici ne comporte pas le sujet) ; par convention, le sujet occupe la position 0, les compléments la position 1, 2 ou 3 (il est considéré qu'il y a au maximum 3 compléments) ;
- la nature prépositionnelle ou non des compléments ; pour les compléments prépositionnels, sont distingués ceux qui sont introduits par les préposition *à*, *de*, *avec*, *Loc*, et autres prépositions. Pour les prépositions *à* et *de*, le critère utilisé est purement surfacique. Pour la préposition *avec*, le critère utilisé est celui des constructions symétriques, voir ci-dessous. Pour la préposition *Loc*, le

critère dépend de la sémantique locative du verbe, critère décrit en détail dans (Guillet et Leclère 1992). Le terme *Loc* regroupe des prépositions variées : voir à *Paris* versus *en Avignon* où le choix de la préposition dépend du nom qu'elle introduit à rapport locatif égal, et *sur la table* versus *sous la table* où le choix de la préposition dépend du rapport locatif ;

- la nature de la réalisation du sujet et des compléments ; en particulier sont distinguées les réalisations sous forme de complétive (notée **Que P**), d'infinitive (notée **V-inf**) et de syntagme nominal (la valeur par défaut, notée **N**).

A titre d'illustration, la propriété définitoire de la Table 9 est  $N_0 V (Que P)_1$  à  $N_2$  : cette table regroupe des verbes comme *dire*, *dissimuler* et *ordonner*, dont le cadre de sous-catégorisation peut se caractériser par une complétive objet et un complément nominal introduit par la préposition *à* (*Luc a dit/dissimulé/ordonné à Marie que Zoé chante*)<sup>3</sup>.

Une propriété définitoire peut aussi indiquer qu'un élément de la table entre dans deux constructions qui sont généralement reliées par un lien de paraphrase. Ainsi la table 35S regroupe les verbes intransitifs « symétriques » qui se caractérisent par deux constructions,  $N_0 V$  avec  $N_1$ , et  $N_0$  et  $N_1 V$ , voir *Luc flirte avec Zoé* et *Luc et Zoé flirtent (ensemble)*.

Enfin de nombreuses propriétés définitoires incluent des informations sémantiques élémentaires. Par exemple, des informations sur les classes des noms têtes des syntagmes nominaux (humain, concret, pluriel, etc.). Ou encore, des informations sur la sémantique des procès : ainsi les verbes entrant dans la construction  $N_0 V N_1$  de  $N_2$  ont été divisés en deux paquets : la table 37E regroupe les procès d'enlèvement (*Luc a débarassé le grenier de ses caisses* signifie que *Luc a enlevé les caisses du grenier*), tandis que les tables 37M (tables 37M1, 37M2, ..., 37M6) regroupent les procès d'ajout (*Luc a muni la porte d'un verrou* signifie que *Luc a mis un verrou sur la porte*). Les tables 37Mi se distinguent par des propriétés très diverses (morphologiques, sémantiques ou autres) qui ne relèvent pas du cadre de sous-catégorisation et qui auraient pu/dû figurer en colonne dans une unique table 37M. La raison de ce découpage est principalement numérique : la table 37M aurait regroupé 890 verbes, et il a été considéré que la consultation manuelle d'une matrice de 890 lignes était difficile, d'où sa division en six sous-tables. Ce critère numérique n'est pas pertinent pour le Traitement Automatique des Langues, néanmoins il nous offre une belle transition pour aborder l'aspect matriciel des tables.

Les tables se présentent donc sous forme de matrice de + et - où les colonnes indiquent les propriétés qui varient d'un élément à l'autre. Ainsi, dans la Table 9, une colonne intitulée « **de V<sup>2</sup> W** » permet de coder si un verbe appartenant à cette Table autorise que son complément direct (de position 1) soit une infinitive introduite par *de* et

<sup>3</sup> Dans la propriété définitoire de la Table 9, le complément indirect apparaît après la complétive. Il n'empêche que cet ordre peut être inversé dans une phrase respectant la propriété définitoire.

contrôlée par  $N_2$  (*Luc a ordonné/dit à Zoé de chanter*, versus *\*Luc a dissimulé à Zoé de chanter*). Une autre colonne intitulée «  $Aux V^0 W$  » permet de coder si un verbe appartenant à la Table 9 autorise que son complément direct soit une infinitive directe à un temps composé et contrôlée par  $N_0$ <sup>4</sup> (*Luc a dit/dissimulé à Zoé avoir chanté*, versus *\*Luc a ordonné à Zoé avoir chanté*). En fait, la situation est plus compliquée car il peut y avoir une structure hiérarchique entre colonnes. Ainsi la colonne «  $de V^2 W$  » dépend d'une colonne qui indique que la complétive est au subjonctif (*Luc a dit/ordonné à Léa que Zoé parte demain*) tandis que la colonne «  $Aux V^0 W$  » dépend d'une colonne qui indique que la complétive est à l'indicatif (*Luc a dit/dissimulé à Léa que Zoé part demain*). De ce fait, pour convertir en un format tel que celui du *Lefff* les informations codées dans la Table 9, il faut comprendre que cette table regroupe (au moins) deux ensembles de verbes : l'un composé de verbes, comme *ordonner*, dont la complétive est au subjonctif et qui permettent une infinitive en  $de V^2 W$ , l'autre, comme *dissimuler*, composé de verbes dont la complétive est à l'indicatif<sup>5</sup> et qui permettent une infinitive en  $Aux V^0 W$ . Le verbe *dire* appartient à ces deux ensembles.

Ces dépendances complexes entre colonnes, c'est-à-dire entre propriétés syntaxiques, ont d'ailleurs été modélisées (manuellement) pour certaines tables par (Gardent *et al.* 2006), sous forme de graphes. Ces graphes, qui modélisent également les nombreuses informations implicites dans les tables, sont destinées à l'extraction automatique d'un lexique TAL, nommé SynLex, à partir des tables traitées<sup>6</sup>.

Nous terminons cette section sur les colonnes en indiquant que diverses colonnes permettent d'indiquer les propriétés de pronominalisation et de cliticisation des différents compléments. Rappelons que ces propriétés peuvent être considérées comme des propriétés définitoires dans le lexique Proton, maintenant DICOVALENCE (van den Eynde et Mertens 2006).

## 2.2. Les constructions impersonnelles et l'outil ILIMP

Nous nous intéressons ici aux phrases dont le sujet est le pronom impersonnel *il* (ce pronom est aussi appelé pléonastique ou explétif). Comme la plupart des phénomènes linguistiques, les constructions impersonnelles reposent sur des conditions tant lexicales que syntaxiques. Par exemple, l'adjectif *violet* ne peut jamais être la tête lexicale d'une phrase impersonnelle, (1a), l'adjectif *probable* ancre une phrase impersonnelle lorsqu'il est suivi d'un complément phrastique, (1b), l'adjectif *difficile* ancre une phrase impersonnelle (resp. personnelle) lorsqu'il est suivi d'une infinitive introduite par la préposition *de* (resp. *à*), (1c) et (1d).

(1)a Il est violet

<sup>4</sup> Une autre colonne intitulée «  $V^0 W$  » permet de coder des phrases comme *Luc dit/prétend être le Messie* qui sont plus naturelles sans complément de type à  $N_2$ .

<sup>5</sup> Sans parler de l'induction du subjonctif en mode non-assertif.

<sup>6</sup> On pourra se reporter à la page Internet du projet : <http://libresource.inria.fr/projects/SynLex>

- b Il est probable que Fred viendra
- c Il est difficile **de** résoudre ce problème
- d Il est difficile à résoudre [ce problème]<sup>7</sup>

De ce fait, le lexique-grammaire du français est une ressource linguistique appropriée pour répertorier l'ensemble des constructions impersonnelles. Ce travail a été effectué lors de la réalisation d'ILIMP (Danlos 2005). ILIMP est un outil qui prend en entrée un texte brut (sans annotation linguistique) rédigé en français et qui fournit en sortie le texte d'entrée où chaque occurrence du pronom *il* est décorée de la balise [ANA] pour anaphorique ou [IMP] pour impersonnel. Cet outil a été conçu en vue de la résolution des anaphores : il permet de distinguer les occurrences anaphoriques du pronom *il*, pour lesquelles un système de résolution des anaphores doit chercher un antécédent, des occurrences où *il* est un pronom impersonnel pour lequel la recherche d'antécédent ne fait pas sens. Nous allons voir comment il peut être utilisé dans un lexique syntaxique comme le *Lefff*. Auparavant, présentons-le brièvement.

### 2.2.1. ILIMP : aspects informatiques

Pour l'aspect informatique, ILIMP repose sur UNITEX<sup>8</sup> qui est un logiciel permettant d'écrire des patrons linguistiques (expressions régulières ou automates récursifs) qui sont localisés dans le texte d'entrée, avec un éventuel ajout d'annotations lorsque les automates sont en fait des transducteurs. Pour ILIMP, l'idée de base consiste à écrire (manuellement) un ensemble de transducteurs comme celui présenté en (2) sous une forme linéaire simplifiée. La balise [IMP] est l'ajout d'information amenée par l'aspect transducteur de (2). Les éléments entre chevrons de (2) se glosent de la façon suivante : <être.V:3s> correspond à toutes les formes du verbe *être* conjugué à la troisième personne du singulier, <Adj1.ms> correspond aux adjectifs masculins singuliers de la classe Adj1 qui regroupe des adjectifs se comportant comme *difficile*, <V:W> correspond aux verbes à l'infinitif.

(2) Il [IMP] <être.V:3s> <Adj1.ms> de <V:W>

La balise [IMP] vient décorer les occurrences de *il* qui apparaissent dans les phrases correspondant au patron de (2). Cette balise vient donc décorer *il* dans (1c). La balise [ANA] est la balise par défaut : elle vient décorer les occurrences de *il* qui n'ont pas été balisées par [IMP]. Cette balise vient décorer *il* dans (1d). Néanmoins, la situation est un peu plus complexe, car il existe une troisième balise [AMB] — abréviation de « ambigu » — qui sera expliquée ci-dessous.

### 2.2.2. Aspects linguistiques : les différentes constructions impersonnelles

Pour l'aspect linguistique, ILIMP repose sur le lexique-grammaire. Plus précisément, nous avons extrait manuellement du lexique-grammaire tous les items lexicaux qui

<sup>7</sup> Un GN entre crochets indique un antécédent possible du sujet anaphorique *il*.

<sup>8</sup> UNITEX est un logiciel sous licence GPL, dont l'ancêtre est INTEX. La documentation et le téléchargement de UNITEX se trouvent sur le site <http://ladl.univ-mlv.fr>.

peuvent ancrer une phrase impersonnelle en enregistrant leur complémentation syntaxique. Présentons un bref aperçu des constructions impersonnelles du français que nous avons recensées.

On peut distinguer les constructions intrinsèquement impersonnelles, qui ne peuvent avoir comme sujet que *il*, des constructions avec un « sujet profond extraposé ».

Parmi les premières, on trouve 45 verbes météorologiques de la table 31I de (Boons *et al.*, 1976a) (*Il pleut, Il vente*), 21 verbes de la table 17 de (Gross, 1975) (*Il faut du pain /que Fred vienne*) et 38 expressions figées de (Gross, 1993) (*Il était une fois, quoi qu'il en soit*), plus quelques verbes des tables 7 et 8 (voir ci-dessous).

Pour les constructions impersonnelles à sujet profond extraposé, on peut distinguer celles à sujet phrastique de celles à sujet uniquement nominal.

Parmi les premières, on trouve 682 adjectifs dispersés dans les tables de (Picabia, 1978) et (Meunier, 1981) (*Il est probable que Fred viendra*), 88 expressions être Prép X des Tables Z5P et Z5D de (Danlos 1980) (*Il est de règle de porter un chapeau*), 21 verbes de la Table 5 de (Gross 1975) (*Il plaît à Zoé que Luc vienne*), et enfin 140 verbes de la Table 6 et 92 verbes de la Table 9 de (Gross 1975) construits au passif ou au *se-moyen* (*Il a été dit/se raconte que Fred viendra*).

Les constructions impersonnelles à sujet extraposé nominal ont pour tête lexicale des verbes qui sont dispersés dans les tables élaborées par (Boons *et al.* 1976b). On peut distinguer d'un côté des verbes comme *manquer* ou *rester* dont l'emploi en construction impersonnelle est tout à fait courant (*Il manque /reste du pain*), et de l'autre côté des verbes « inaccusatifs » (*Il est venu trois personnes*) ou des verbes construits au passif (*Il a été mangé trois gâteaux*), dont l'emploi dans une construction impersonnelle relève d'un niveau de langue châtié. Seuls les verbes du type *manquer* ou *rester* ont été recensés. Pour ces verbes, le statut impersonnel ou non du sujet dépend du déterminant introduisant le GN sujet extraposé, voir la paire en (3), ou du nom tête de ce GN, voir la paire en (4).

(3)a Il manque **du** poivre (dans cette maison)

b Il manque **de** poivre [ce rôti]

(4)a Il reste la **valise** du chef (dans la voiture)

b Il reste la **priorité** du chef [le chômage]

Dans un très petit nombre de cas (une dizaine), un item lexical peut ancrer une construction impersonnelle ou personnelle avec le même cadre de sous-catégorisation. C'est le cas pour l'adjectif *certain* construit avec un complément phrastique, comme illustré dans la phrase en (5a). Comme les deux lectures de (5a) semblent également fréquentes, *il* dans le patron (5b) reçoit la balise [AMB]. On notera que le cas de *certain* (5a,b), pour lequel on peut considérer qu'il y a un lien de polysémie entre les deux constructions, est différent de celui d'un adjectif comme *digne*, dont les deux constructions (5c) et (5d) correspondent à des entrées lexicales distinctes (homonymiques).



- (5)a Il est certain Fred que viendra (*Jean/Cela est certain que Fred viendra*)  
 b Il [AMB] est certain que P  
 c Il [AMB] est digne de figurer au palmarès [Luc]  
 d Il [AMB] est digne de se comporter de la sorte [*impersonnel*]

Soulignons bien que l'extraction des constructions impersonnelles à partir du lexique-grammaire a été manuelle, et non-automatique comme dans le travail présenté dans (Gardent *et al.* 2006). Cette différence est de taille, comme illustré dans l'exemple qui suit : le cas des verbes à sujet phrastique extraposable qui proviennent des tables de (Gross 1975). Une extraction automatique consisterait à prendre :

- toutes les entrées de la table 5, dont la propriété définitoire est justement l'existence de la construction  $(\text{Que } P)_0 \text{ V Prép } N_1$  associée à la construction impersonnelle  $\text{Il V Prép } N_1 \text{ Que } P$  (voir *Que Luc parte plaît à Zoé* versus *Il plaît à Zoé que Luc parte*),
- toutes les entrées de la table 17, dont la propriété définitoire est *Il V (Prép ce) Que P Prép  $N_2$*  (voir *Il semble à Luc que Zoé est partie*),
- les quelques entrées des Tables 7 ( $N_0 \text{ V à ce que } P$ ) et 8 ( $N_0 \text{ V de ce que } P$ ) où il existe une colonne intitulée [extrap] (qui dépend de la colonne intitulée **Sujet**) qui code la possibilité d'extraposer un sujet phrastique (voir *Que Zoé soit partie découle de ce que Luc est arrivé* versus *Il découle de ce que Luc est arrivé que Zoé soit partie*).

Une telle extraction automatique aurait induit du bruit et du silence. Commençons par le bruit. La table 5 contient un verbe comme *galoper* dont l'entrée dans la Table 5 est justifiée par l'exemple en (6) (avec *Prép* = Loc).

(6) Il a galopé dans l'esprit de Luc que Zoé devait tondre

L'exemple (6) est clairement métaphorique. De ce fait, nous ne voulons pas compter *galoper* comme tête lexicale d'une possible construction impersonnelle. La raison est que nous ne voulons pas multiplier **inutilement** l'ambiguïté dans les traitements automatiques. Que ce soit dans ILIMP ou dans les analyseurs syntaxiques qui reposent sur le Lefff, considérer *galoper* comme tête lexicale d'une possible construction impersonnelle amènerait à considérer (6) et (7) ci-dessous comme ambiguës entre une lecture impersonnelle et personnelle, alors qu'elles ne le sont nullement. Dans la lecture impersonnelle de (6) ou (7), l'objet direct de *tondre* n'est pas réalisé. Dans la lecture personnelle, il est réalisé sous forme de pronom relatif *que*, la phrase ayant la structure  $N_0 \text{ galoper Loc } N_1$  (table 35L).

(7) a Il a galopé dans le champ de Luc que Zoé devait tondre [il = le cheval]

Certes, on peut arguer que les exemples (6) et (7) se différencient par les noms *esprit* et *champ*, l'un à caractère abstrait (*esprit*) l'autre à caractère concret (*champ*), et que cette différence peut permettre de désambiguïser ces phrases. Néanmoins, il est bien connu que ces traits sémantiques sont difficilement codables. Aussi, nous pensons que nous ne pourrions pas désambiguïser (6) ou (7) par manque d'informations

sémantiques, et nous prenons le parti de ne pas considérer que *galoper* a une entrée dans la table 5, entrée qui serait à la fois rare et métaphorique. Ceci pour ne pas ajouter une ambiguïté artificielle dans les analyseurs, qui doivent déjà faire le départ entre un nombre d'analyses exponentiel en la longueur de la phrase.

Passons au silence. La possibilité d'avoir un sujet phrastique extraposable dans une construction passive ou pronominale n'est pas codée dans les tables à complétive. Pourtant, elle varie d'un verbe à l'autre, y compris dans la même table. A titre d'illustration, considérons la table 6 de propriété définitoire  $N_O V (Que P)_1$ , dont font partie les verbes *spéculer* et *sentir* (*Luc spéculé/sent que Zoé va partir*). Le verbe *spéculer* autorise une construction impersonnelle passive, mais pas *sentir*, voir (8).

(8) a Il a été spéculé que Zoé partira

b \*Il a été senti que Zoé partira<sup>9</sup>

Une colonne intitulée [extrap-passif] et une autre colonne [extrap-pronominale] auraient donc dûes être incluses dans la table 6 et dans d'autres tables. En l'absence d'un tel codage, seul un travail manuel de linguiste permet de rattraper des constructions impersonnelles non codées dans le lexique-grammaire.

C'est donc à partir des bases linguistiques de ILIMP (et non directement à partir de celles du lexique-grammaire) qu'ont été renseignées les constructions impersonnelles dans le *Lefff*, en suivant la modélisation décrite dans la Section 4. 1.

Comme nous allons le voir, le même constat s'applique pour les expressions verbales figées : leur recensement effectué dans le cadre du lexique-grammaire constitue une base d'une valeur inestimable, mais qui demande néanmoins un travail linguistique conséquent pour pouvoir être intégré dans un lexique syntaxique comme le *Lefff*.

### 2.3. Expressions verbales figées

M. Gross a réalisé un lexique-grammaire de 44 000 expressions verbales figées, qu'il n'a malheureusement pas eu le temps de documenter exhaustivement avant sa disparition ; on peut lire cependant (Gross 1982, Gross 1993). Ce lexique-grammaire suit les mêmes principes que ceux établis pour les constructions verbales libres (voir section 2.1) mais il comporte aussi d'autres critères, en particulier sur la structure interne des expressions figées. Nous allons décrire cela plus précisément.

Une expression verbale figée est telle qu'une des positions au moins est réalisée sous forme d'un syntagme nominal dont la tête est une constante notée *C*. Rappelons que, pour les expressions verbales libres, les positions sont réalisées sous forme de complétive, d'infinitive ou de syntagme nominal libre (i.e. dont la tête peut varier librement à l'intérieur d'une classe sémantique). Les propriétés définitoires des tables de figées définissent des cadres de sous-catégorisation, comme c'est le cas pour les

---

<sup>9</sup> L'acceptabilité de cet exemple semble augmenter si l'on rajoute un élément en *par* qui introduit un syntagme nominal indéfini ou collectif (*il a été senti [par un grand nombre de personnes / par tout l'assistance / ??par Luc] que Marie partira*).

tables de non-figées. A titre d'illustration la structure  $(\text{Que } P)_0 \text{ V Prép } N_1$  peut se réaliser de deux façons différentes :

- si  $N_1$  est un syntagme nominal libre, alors le verbe  $V$  appartient à la table 5, ce qui est le cas de *plaire* (*Que P plaît à Zoé* — cas mentionné en 2.2) ;
- si  $N_1$  est une constante alors on a affaire à une expression verbale figée de la table C5, comme illustré en (9) :

(9a) *Que P coule de (source)*<sub>1</sub>

b *Que P reste sur (l'estomac de Zoé)*<sub>1</sub>

La table C5 comporte donc au moins trois types d'éléments lexicaux pour chaque entrée : un verbe, une préposition et une constante (alors que la table 5 n'en comporte que deux : le verbe et la préposition). Néanmoins, on observe une différence entre les exemples (9a) et (9b) : la constante en (9a) est un nom nu (*source*) tandis que celle de (9b) (*estomac*) est la tête d'un GN complexe qui est composé d'un déterminant (*le*) suivi du nom tête constant (*estomac*), ce nom étant lui-même suivi d'un complément de nom introduit par la préposition *de* qui introduit un GN libre (obligatoirement humain). La table C5 comporte donc plusieurs éléments lexicaux permettant de décrire la structure interne du GN dont le nom tête est une constante.

La table C5 regroupe donc toutes les expressions verbales figées respectant la propriété définitoire quelles que soient leur structure interne. Par contre, certaines autres tables prennent en compte la structure interne dans leur propriété définitoire. C'est le cas pour les tables notées  $N_0 \text{ V } C_1$  qui se divisent en plusieurs sous-tables :

- la table C1i lorsque la constante est introduite par un indéfini, voir *Luc amorce (un virage)*<sub>1</sub>,
- la table C1g lorsque la constante est introduite par un possessif référant au sujet, voir *Luc a cassé (sa<sup>0</sup> tirelire)*<sub>1</sub>,
- la table C1dN lorsque la constante est suivie d'un complément de nom introduit par la préposition *de*, voir *Luc boit (les paroles de Zoé)*<sub>1</sub>,
- etc<sup>10</sup>...

Enfin les propriétés définitoires des tables de figées peuvent comporter des critères divers, dont le fait que le verbe est *être* ou *avoir*.

Il est connu que la notion d'expression figée est difficile à cerner (Lamiroy et Klein 2006) : on observe un continuum entre expressions libres et expressions totalement figées, ces dernières étant caractérisables par une totale opacité sémantique et une absence de variation syntagmatique ou paradigmatique. Entre ces deux extrêmes, on observe des degrés variables de figement. La politique de M. Gross a été de considérer comme figée toute expression qui s'inscrit dans un paradigme contraint. Cette vision

<sup>10</sup> Nous ne pouvons décrire avec précision l'inventaire exact des tables en  $N_0 \text{ V } C_1$  car actuellement (fin octobre 2006) seule environ la moitié des tables sont distribuées à la communauté.

large de la notion de figée a débouché sur 44 000 expressions verbales, dont le caractère figé est souvent contestable. Ainsi le projet BFQS<sup>11</sup>, qui travaille sur les variantes belges, françaises, québécoises et suisses des expressions figées est parti du recensement de M. Gross mais n'a retenu qu'une partie de ces expressions (Lamiroy *et al.* 2003).

Pour des raisons parallèles, l'utilisation du lexique-grammaire des figées pour étendre un lexique syntaxique comme le *Lefff* demande de faire un gros travail, entre autres pour déterminer ce qu'il est souhaitable de retenir comme figé en vue d'une analyse syntaxique. Ce travail n'ayant pas encore été mené à son terme, contrairement au travail linguistique effectué pour les constructions impersonnelles, nous ne présenterons en section 4.2 qu'une modélisation préliminaire des expressions verbales figées, réalisée seulement sur quelques tables.

### 3. Le *Lefff*

Le *Lefff*, Lexique des Formes Fléchies du Français (Sagot *et al.* 2006), est un lexique électronique de la langue française, librement disponible<sup>12</sup>. Il associe à chaque forme fléchie des informations morphologiques (lemme, étiquette morphologique) et syntaxiques (cadre de sous-catégorisation, autres propriétés syntaxiques). Il est destiné à être directement utilisé dans des applications de Traitement Automatique des Langues, en essayant toutefois d'être indépendant des choix théoriques de ses utilisateurs, et en particulier des théories syntaxiques.

Le *Lefff* a la particularité de chercher un équilibre entre la pertinence de la modélisation linguistique et l'adéquation aux besoins opérationnels. Ceci se traduit dans la façon dont l'information linguistique est représentée, mais également dans la façon dont elle est ajoutée au *Lefff*. En effet, en plus de méthodes manuelles, des techniques d'acquisition automatique d'information ont été utilisées pour compléter et corriger le *Lefff*. Cependant, ces techniques sont toujours suivies d'étapes de validation manuelle, pour permettre la préservation d'un niveau satisfaisant de qualité.

Le *Lefff* est une ressource à large couverture : il rassemble plus de 110 000 lemmes, auxquels correspondent plus de 520 000 entrées<sup>13</sup>. Parmi ces entrées, les formes verbales, un certain nombre de formes nominales et adjectivales (mais pas toutes), ainsi que les prépositions et d'autres types d'entrées (y compris des constructions à

---

<sup>11</sup> On pourra se reporter sur Internet à la page d'accueil du projet, à l'adresse suivante : <http://bfqs.fltr.ucl.ac.be>

<sup>12</sup> On se référera au site Internet du *Lefff*, à l'adresse suivante : [www.lefff.net](http://www.lefff.net)

<sup>13</sup> De plus, certaines de ces entrées sont encore factorisées : par exemple, pour un verbe du premier groupe, les première et troisième personne du singulier du présent de l'indicatif et du subjonctif sont regroupés en une seule entrée factorisée, avec une étiquette morphologique incluant deux disjonctions

verbe support) sont associées à des cadres de sous-catégorisation spécifiques<sup>14</sup>, ainsi qu'à des informations syntaxiques complémentaires (contrôle, attributif, contraintes sur les syntagmes régis,...).

### 3.1. Le développement du *Lefff* : historique et architecture

#### 3.1.1. *Le Lefff 1, lexique morphologique empirique des verbes du français*

Le développement du *Lefff* a commencé en 2003, à partir du constat suivant : à cette époque, il n'existait pas de lexique syntaxique pour le français qui soit librement utilisable et dont la couverture soit importante. Le projet a donc été lancé de construire un tel lexique, avec le double objectif qu'il soit adapté au TAL tout en restant linguistiquement pertinent.

La première étape dans le développement d'un lexique syntaxique est celui du lexique morphologique sous-jacent. Dans (Clément, Sagot et Lang, 2004), les auteurs décrivent la première version d'une technique d'acquisition automatique de lexique morphologique, à partir d'un corpus brut et d'une description morphologique de la langue étudiée. Cette technique, dont une version plus aboutie est présentée dans (Sagot, 2005), repose sur la variabilité morphologique. Pour cette raison, elle n'a été appliquée pour le français que pour l'acquisition d'un lexique verbal. C'est ainsi qu'a été mis à disposition de la communauté le *Lefff 1*, lexique morphologique verbal du français, acquis automatiquement et validé manuellement.

#### 3.1.2. *Le Lefff 2 : introduction des informations syntaxiques*

Par la suite, des cadres de sous-catégorisation ont été ajoutés à ce lexique de formes verbales. Pour certains verbes, une ébauche de structure syntaxique avait été développée par Lionel Clément dans des travaux non publiés. Ces informations relativement parcellaires ont servi de base à la constitution d'un graphe de classes syntaxiques, définies par héritage de propriétés syntaxique atomiques, propriétés elles-mêmes définies de façon indépendante de la définition des classes. Ces classes permettent de définir un **lexique intensionnel** formé de triplets (*lemme, classe morphologique, classe syntaxique*). La classe morphologique, grâce à un outil de flexion généré automatiquement à partir d'une description morphologique du français développée pour le *Lefff*, permet de construire le lexique morphologique extensionnel, qui associe à chaque forme fléchie son lemme, son étiquette morphologique, et une propriété morphosyntaxique (*Default, Infinitif, ParticipePasséActif, ParticipePasséPassif*) qui influera sur sa structure syntaxique.

Une deuxième phase permet, à partir de la classe syntaxique associé au lemme et de cette propriété morphosyntaxique, de construire le **lexique extensionnel** complet, qui associe à chaque forme fléchie une catégorie (ou partie du discours), éventuellement

---

<sup>14</sup> Par opposition aux cadres de sous-catégorisation génériques (et très tolérants) attribués pour l'instant aux noms et aux adjectifs pour lesquels le travail descriptif reste à faire (ou à extraire d'autres ressources).

un poids (calculé selon des heuristiques ou renseigné manuellement), et une structure syntaxique complète, y compris un cadre de sous-catégorisation.

Dans un premier temps, un lexique syntaxique extensionnel des formes verbales du français a donc été constitué et mis à disposition de la communauté : il s'agit de la première version du *Lefff* 2.0. Dans cette version, les cadres de sous-catégorisation étaient exprimés dans un formalisme très proche du formalisme LFG, quoiqu'avec un ensemble de « fonctions grammaticales » plus large et plus proche de la syntaxe qu'en LFG standard.

### *3.1.3. Le Lefff 2.1 puis 2.2 : prise en compte de toutes les catégories*

En parallèle à ces développements, le développement d'un lexique syntaxique couvrant toutes les catégories était en cours. Alors que les catégories fermées ont été renseignées principalement à la main, les autres catégories ouvertes ont été constituées par divers moyens complémentaires :

- acquisition automatique (avec validation manuelle) à l'aide de techniques déjà citées (Clément, Sagot et Lang 2004, Sagot 2005) ;
- acquisition automatique (avec validation manuelle) d'informations syntaxiques atomiques (*cf.* Sagot 2006 : ch. 7) ;
- pour certains noms, adjectifs et adverbes, exploitation du lexique morphologique Multext pour le français (Véronis 1998), dont la libre exploitation nous a été autorisée explicitement par son principal auteur ;
- corrections et ajouts manuels ou guidés par des techniques automatiques, comme par exemple la fouille d'erreurs dans les sorties d'analyseurs syntaxiques (Sagot et de La Clergerie 2006).

Le tableau 1 montre le jeu de catégories utilisées aujourd'hui dans le *Lefff*. Il a évolué avec le temps, au fur et à mesure que l'utilisation du *Lefff* dans des analyseurs a révélé la nécessité de distinguer certaines catégories nouvelles. On notera cependant que les utilisateurs du *Lefff* utilisent presque toujours un convertisseur, qui leur permet d'adapter le *Lefff* à leurs besoins, par exemple en extrayant de la structure syntaxique des informations permettant d'augmenter encore le nombre de catégories.

En parallèle, le format de représentation des cadres de sous-catégorisation a évolué. Sous l'influence de réflexions théoriques, de constats pratiques (en particulier lors de l'utilisation du *Lefff* dans divers systèmes d'analyse), et de travaux de comparaison avec d'autres ressources lexicales apparaissant progressivement (SynLex), ce format de représentation des données syntaxiques a évolué. Le changement le plus profond a été le découplage entre les fonctions syntaxiques sous-catégorisées et les réalisations possibles de ces fonctions syntaxiques. Une description complète de ce modèle fait l'objet du paragraphe suivant.

adj	adjectif ( <i>beau</i> )	coo	coordonnant ( <i>et,...</i> )
adjPref	préfixe adjectival ( <i>ex- dans ex-ministre</i> )	csu	subordonnant ( <i>parce que, si...</i> )
adv	adverbe ( <i>lentement</i> )	det	déterminant ( <i>un, le,...</i> )
advPref	préfixe adverbial ( <i>sur- dans surpuissant</i> )	epsilon	forme à ne pas prendre en compte (les guillemets sont par exemple ambigus entre epsilon et ponctw)
advneg	pas, point, guère, etc.	etr	séquence en langue étrangère (entité nommée _ETR)
auxAvoir	formes de l'auxiliaire <i>avoir</i>	ilimp	<i>il</i> impersonnel
auxEtre	formes de l'auxiliaire <i>être</i>	nc	nom commun ( <i>pomme</i> )
caimp	<i>ça</i> impersonnel	ncpred	nom commun prédicatif ( <i>conscience dans prendre conscience,...</i> )
ce	ce dans à/de ce que	np	nom propre (Nancy, Air France)
cf	complément figé standard	poncts	ponctuation forte (point final,...)
cfi	complément figé inséparable de sa tête syntaxique (à certains modifieurs près)	ponctw	ponctuation faible (virgule,...)
cla	clitique accusatif ( <i>le, la, les,...</i> )	prel	pronom relatif ( <i>qui, que, dont,...</i> )
clar	clitique accusatif réflexif ( <i>se</i> )	prep	préposition ( <i>dans, au moyen de</i> )
cld	clitique datif ( <i>lui, leur,...</i> )	pres	« présentatif » (interjections, <i>au revoir,...</i> )
cldr	clitique datif réflexif ( <i>se</i> )	pri	pronom interrogatif ( <i>qui, où,...</i> )
clg	clitique « génitif » ( <i>en</i> )	pro	pronom plein ( <i>moi, l'un, une,...</i> )
cll	clitique « (dé)locatif » ( <i>en, y</i> )	que	que dans à/de ce que
cln	clitique nominatif ( <i>il,...</i> )	sbound	frontière de phrase non lexicalisée (entité nommée _SENT_BOUND)
clneg	clitique négatif ( <i>ne</i> )	suffAdj	suffixe adjectival ( <i>-né dans artiste-né</i> )
clr	clitique réflexif non anaphorique ( <i>se</i> )	v	forme verbale ( <i>manger, mange, mangé</i> )

Tableau 1. Jeu de catégories utilisées dans le Lefff version 2.2.2

C'est donc aujourd'hui un lexique syntaxique à large couverture pour le français, qui ne se restreint ni aux seules formes verbales ni aux seules informations morphologiques, qui est mis à disposition. Le Lefff, actuellement<sup>15</sup> en version 2.2.2, est entièrement téléchargeable sous sa forme extensionnelle, et sera prochainement téléchargeable également sous sa forme intensionnelle, sous une licence libre (LGPL-LR), sur le site internet [www.lefff.net](http://www.lefff.net).

<sup>15</sup> Fin octobre 2006

### 3.2. Modélisation des informations syntaxiques

Une entrée simple du *Lefff* ressemble à ce qui suit :

mange v [pred='manger<Suj:sn|cln,Obj:(sn|cla)>', cat=v, @PS13s]

On distingue bien la catégorie et la structure syntaxique. Cette dernière comporte un « pred » à la LFG, composé d'un identifiant sémantique, souvent identique au lemme, et d'un cadre de sous-catégorisation que nous étudions en détails ci-dessous. Ici, il s'agit d'un sujet nominal ou clitique obligatoire et d'un objet nominal ou clitique facultatif (comme indiqué par les parenthèses). Enfin, la catégorie morphosyntaxique est indiquée par l'attribut *cat*<sup>16</sup>, ainsi qu'une macro résumant l'étiquette morphologique<sup>17</sup> (les macros sont introduites par le symbole « @ »).

Par rapport au format LFG du *Lefff* 2.0, le découplage entre fonctions syntaxiques et réalisations résoud un grand nombre de difficultés théoriques et pratiques<sup>18</sup>. Pour n'en citer qu'une, ceci permet de représenter de façon satisfaisante des cadres de sous-catégorisation où deux « fonctions grammaticales » identiques pouvaient coexister, et en particulier deux compléments indirects ou obliques introduits par la même préposition (il en est ainsi, par exemple, de tous les verbes transitifs passivables sous-catégorisant un complément en *par* dès la forme active, tels que *diviser*, *multiplier*,... : *Le taux de chômage a été divisé par deux par les dernières réformes*).

#### 3.2.1. Le cadre de sous-catégorisation : les fonctions syntaxiques et leurs réalisations

Un cadre de sous-catégorisation est constitué d'une liste (éventuellement vide) de *fonctions syntaxiques*, chacune d'entre elle se voyant attribuer un certain nombre de *réalisations* (de surface) possibles. Ces réalisations peuvent être des syntagmes et des clitiques. Toutefois, une fonction syntaxique peut n'être que facultativement réalisée. La position d'une fonction dans la liste de fonctions que constitue un cadre de sous-catégorisation est le *rang* de cette fonction dans ce cadre<sup>19</sup>. Le format utilisé pour représenter ces informations est décrit dans la grammaire donnée au tableau 2.

Nous avons pris en compte les arguments et les conclusions de divers auteurs (Lionel Clément, Claire Gardent, Éric de La Clergerie, Éric Laporte, Sylvain Kahane, Susanne Salmon-Alt), et plus particulièrement les travaux de Karel van den Eynde et Piet Mertens pour le lexique de valence Proton, aujourd'hui DICOVALENCE, développé dans

<sup>16</sup> Cette dernière est souvent identique à la catégorie proprement dite (et qui est définie sur des critères syntaxiques), mais pas toujours. Ainsi, le préfixe *ex-* a une catégorie *adjPref*, mais sa structure syntaxique comporte *cat=adj*.

<sup>17</sup> Cette étiquette, comme indiqué dans une note précédente, est ici à double disjonction (personne 1 et 3, temps P=présent de l'indicatif et S=présent du subjonctif)

<sup>18</sup> Ce découplage correspond à deux dimensions différentes des métagrammaires de Candito, à la distinction entre paradigmes et réalisations dans le lexique DicoValence, et à la distinction entre et dans l'utilisation de la norme LMF pour le français proposée par (Alt, Danlos, Laporte, Sagot, *non publié*).

<sup>19</sup> Les rangs 1, 2, 3, 4 correspondent aux positions 0, 1, 2, 3 qui indiquent les noms de compléments dans les cadres syntaxiques définitoires des tables du lexique-grammaire (*cf.* section 2).



Cadre	→	$\epsilon$
Cadre	→	< Fonction_liste >
Fonction_liste	→	Fonction , Fonction_liste
Fonction_liste	→	Fonction
Fonction	→	Nom_de_fonction : Realisation_obligatoire
Fonction	→	Nom_de_fonction : Realisation_facultative
Nom_de_fonction	→	<i>voir ci-dessous</i>
Realisation_facultative	→	( Realisation_liste )
Realisation_obligatoire	→	Realisation_liste
Realisation_liste	→	Realisation   Realisation_liste
Realisation_liste	→	Realisation
Realisation	→	<i>voir ci-dessous</i>

Tableau 2. Grammaire BNF des cadres de sous-catégorisation dans le Lefff 2.2.2.

l'approche pronominale (van den Eynde et Mertens 2006). Ceci nous a conduit à la liste de fonctions syntaxiques ci-dessous, indiquées avec leurs critères définitoires.

**Suj** : Fonction *sujet*. La forme clitique est celle d'un clitique nominatif personnel. À l'actif, elle est réalisée canoniquement en position pré-verbale<sup>20</sup> (avec accord). Elle correspond au paradigme P0 du DICOVALENCE.

**Obj** : Fonction *objet (direct)*. La forme clitique est celle d'un clitique accusatif, ou d'un clitique génitif à sens partitif. Un verbe sous-catégorisant une fonction objet est dit transitif (direct). Si le verbe est passivable<sup>21</sup>, cette fonction est translatée, pour devenir fonction sujet. Elle correspond approximativement au paradigme P1 du DICOVALENCE.

**Objà** : Fonction *objet indirect introduit par à*, ou fonction *à-objet*. Est substituable un syntagme prépositionnel de la forme *à* + *pronom non-clitique*<sup>22</sup>. La cliticisation est possible à l'aide du clitique datif (dans tous les cas, ou bien seulement dans le cas humain), peut être possible dans certains cas seulement (non humain) à l'aide du clitique locatif *y*, ou ne pas être possible du tout. Elle se distingue de la fonction locative par la non-substituabilité des pronoms *là*, *ici*, *là-bas*. Elle correspond au paradigme P2 du DICOVALENCE.

**Objde** : Fonction *objet indirect introduit par de*, ou fonction *de-objet*. Est substituable un syntagme prépositionnel de la forme *de* + *pronom non-clitique*. La cliticisation est possible à l'aide du clitique génitif. Elle se distingue de la fonction délocative Dloc par la non-substituabilité avec les locutions pronominales *de là*, *d'ici*. Elle correspond au paradigme P3 du DICOVALENCE.

<sup>20</sup> Ce n'est cependant pas le cas en présence d'un *il* impersonnel avec sujet extraposé.

<sup>21</sup> Ce qui n'est pas automatique, même en cas de transitivité : *Ce problème regarde Marie* / *\*Marie est regardée par ce problème*.

<sup>22</sup> *Pronom non-clitique* est à prendre ici au sens de (van den Eynde et Mertens 2006)

**Loc** : Fonction *locative*. Les pronoms *là*, *ici*, *là-bas* sont substituables. La cliticisation, si elle est possible, se fait à l'aide du clitique locatif *y*. Elle correspond au paradigme PL du DICOVALENCE.

**Dloc** : Fonction *délocative*. Les locutions pronominales *de là*, *d'ici* sont substituables. La cliticisation, si elle est possible, se fait à l'aide du clitique génitif *en*. Elle correspond au paradigme PDL du DICOVALENCE.

**Att** : Fonction *attributive*. Cette fonction, dont les propriétés de cliticisation sont variables, regroupe les attributs du sujet ou d'un des objets (objet, à-objet). Les situations couvertes sont variées : *prendre Pierre [pour Adj / det N]<sub>Att</sub>*, *nommer Pierre [président]<sub>Att</sub>*, *regarder Pierre [courir]<sub>Att</sub>*, *trouver Pierre [Adj]<sub>Att</sub>*, *voir Pierre [(comme) (det) N / (comme) Adj]<sub>Att</sub>*.

**Obl** et **Obl2** : Fonctions *obliques*. Ces fonctions, qu'aucun critère ne distingue l'une de l'autre<sup>23</sup>, abritent les compléments obliques, jamais cliticisables, qui ne rentrent dans aucune des autres fonctions (*y* compris le « complément d'agent » de la grammaire traditionnel dans les constructions passives). Dans un avenir proche, grâce à la disponibilité du lexique DICOVALENCE (ex-Proton), ces fonctions obliques pourront être mieux précisées.

Comme indiqué plus haut, on attribue, dans un cadre de sous-catégorisation donné une disjonction de réalisations à chaque fonction syntaxique présente. Les réalisations possibles sont de trois types :

- Un pronom clitique : clitique nominatif (**cln**), clitique accusatif (**cla**), clitique génitif (**en**), clitique locatif (**y**). On notera que le *se* réfléchi ou réciproque est considéré comme une réalisation de type *cla* ou *cld* selon les cas (*Les époux se disputent / Pierre se laisse cette possibilité*)<sup>24</sup> ;
- Un syntagme direct : syntagme nominal (**sn**), syntagme adjectival (**sa**), syntagme infinitif (**sinf**), syntagme phrastique fini (**scompl**), interrogative indirecte (**qcompl**). Rien n'exclut la possibilité d'introduire également des syntagmes adverbiaux (**sadv**) ;
- Un syntagme prépositionnel : il s'agit d'un syntagme direct précédé d'une préposition, comme **de-sn**, **à-sinf** ou **pour-sa**<sup>25</sup>. Enfin, les notations **à-scompl** et **de-scompl** représentent les réalisations en *à ce que P* et *de ce que P* respectivement.

<sup>23</sup> On n'utilise Obl2 que lorsqu'il y a deux compléments obliques.

<sup>24</sup> Actuellement, toute fonction syntaxique Obj ou Obj<sub>a</sub> réalisable de façon clitique est donc considérée comme pouvant être réalisée par le clitique *se* (réflexif ou réciproque). Ceci est une approximation, que des travaux ultérieurs devront préciser plus avant.

<sup>25</sup> On notera que nous ne distinguons pas les prépositions des complémentisateurs, cette distinction se déduisant de la fonction syntaxique dont on parle. Un **de-sinf** réalisant une fonction objet met en œuvre le complémentisateur *de* (*Jean ordonne à Marie de partir*) ; un **de-sinf** réalisant une fonction de-objet est prépositionnel (*Jean rêve de partir*).

À titre d'exemple, les formes du verbe *ordonner*, utilisées dans une construction personnelle active, seront du type :

ordonnât v [pred='ordonner<sub>1</sub><Suj:sn|cln,Obj:sn|cla|de-sinf|scompl,  
Objà:(à-sn|cld)>',cat=v,@T3s]

On notera que les fonctions syntaxiques dont il s'agit ici sont des fonctions syntaxiques de surface, au sens où un changement de diathèse redistribue certaines fonctions syntaxiques, ainsi que la façon dont elles est réalisées. En revanche, leur rang dans le cadre de sous-catégorisation, qui n'est pas modifié par la diathèse, permet de garder trace de l'identité du rôle sémantique sous-jacent. Ainsi, les deux participes passés (actif et passif) de *manger* ont pour entrées respectives :

mangé v [pred='manger<sub>1</sub><Suj:sn|cln,Obj:(sn|cla)>',cat=v,@active,@avoir,  
@Kms]

mangé v [pred='manger<sub>1</sub><Obl:(par-sn),Suj:(sn|cln)>',cat=v,@passive,@Kms]

Comme on peut le voir, le sujet de la construction active, qui est de rang 1, correspond au complément oblique en **par-sn** de la construction passive, également de rang 1. De même pour l'objet actif qui devient sujet passif. On notera que ces deux participes passés, quoique n'ayant pas le même cadre de sous-catégorisation, sont deux formes issues de la même entrée lexicale au niveau intensionnel. En réalité, une même entrée intensionnelle peut se voir associer une classe syntaxique qui dénote elle-même une disjonction de comportements. C'est ainsi qu'un adjectif comme *envisageable*, une fois les constructions impersonnelles prises en compte (cf. section 4), n'aura qu'une seule entrée au niveau intensionnel, alors qu'au niveau extensionnel, chacune de ses formes aura deux entrées, l'une pour la construction personnelle (*ceci est envisageable/une chose envisageable*), et l'autre pour la construction impersonnelle (*il est envisageable de Vinf/que P*).

### 3.2.2. Autres propriétés syntaxiques

D'autres propriétés syntaxiques complètent le cadre de sous-catégorisation. Pour la plupart des réalisations infinitives (directes ou non) dans les cadres de sous-catégorisation verbaux, une information de contrôle est donnée : selon les cas, le sujet de l'infinitive est « égal » (en un sens qui dépend des théories syntaxiques) au sujet, à l'objet direct ou à l'objet indirect (à-objet) du verbe. On notera que certains verbes sous-catégorisent des fonctions pouvant se réaliser par une infinitive sans qu'il y ait contrôle : *le travail consiste à créer un lexique*. De même, toute fonction attributive à réalisation nominale ou adjectivale possible (directe ou prépositionnelle) est précisée par une indication de la fonction à laquelle l'attribut s'applique (sujet, objet, à-objet).

Une autre propriété (mal) renseignée dans le *Lefff* est relative aux contraintes sur le mode des complétives, qu'elles réalisent des fonctions syntaxiques sous-catégorisées par des entrées verbales, nominales, ou autres. Quatre grands cas de figure sont répertoriés : le cas où les deux modes sont possibles, le mode indicatif obligatoire, le mode subjonctif obligatoire, et le mode dit alternant (subjonctif en cas de non-assertion, indicatif sinon).

Ces propriétés sont indiquées, comme pour l'étiquette morphosyntaxique, par des macros (telles que @CtrlSuj pour le contrôle sujet, @AttObj pour l'attribut de l'objet, @CompSubj pour la complétive au subjonctif). L'idée est que chaque utilisateur du *Lefff* est amené à donner à ces macros la signification appropriée, compte tenu du formalisme ou du contexte d'utilisation de cette information. Toutefois, nous disposons d'une définition de ces macros sous forme de structure de traits avec partage possible, qui représente de façon transparente ce qu'elles veulent dire. Ainsi, on a la définition suivante pour le contrôle sujet :

@CtrlSuj := [ Suj=[1, Obj=[ Suj = [1 ] ]

Ceci indique que le sujet de l'infinitive objet est partagé (par co-indiciation) avec le sujet principal (le mécanisme qui n'applique cette macro que lorsque la réalisation de l'objet est effectivement infinitive n'est pas indiqué) .

Enfin, si les fonctions syntaxiques correspondent à des arguments syntaxico-sémantiques du prédicat concerné (verbal, adjectival, nominal, etc.), certaines entrées s'accompagnent de la présence de clitiques qui ne correspondent à aucun argument syntaxico-sémantique. Outre les pronoms impersonnels *il* et *ça*, que nous traiterons plus bas, c'est le cas du clitique réflexif dans le cas des verbes dits essentiellement pronominaux (*s'évanouir*), du clitique génitif (*en référer à quelque chose*, *en être quelque part*), du clitique locatif (*y passer*), voir de plusieurs d'entre eux (*s'y connaître*, *s'en tirer*). Ces clitiques sont requis à l'aide de macros spécifiques dans la structure syntaxique, dont certaines empruntent une terminologie issue de DICOVALENCE (@pseudo-y, @pseudo-en, @négatif, @pronominal...).

### 3.3. Utilisation en TAL

Le *Lefff* est utilisé par au moins deux systèmes d'analyse très différents pour le français. Le premier d'entre eux est FRMG (Thomasset et de La Clergerie 2005), un analyseur TAG qui repose sur une métagrammaire, laquelle génère une TAG factorisée. Les entrées du *Lefff* sont utilisées comme « hypertags » pour ancrer les quasi-arbres.

Le second analyseur, que nous utiliserons dans la suite de cet article pour illustrer l'analyse de phrases à constructions verbales complexes, est l'analyseur du français construit à l'aide de SxLFG, constructeur d'analyseurs LFG (Boullier et Sagot 2005). Il s'agit d'un analyseur LFG très efficace qui utilise les entrées du *Lefff* comme entrées lexicales LFG.

Ces deux systèmes utilisant le *Lefff* sont utilisés dans différentes expériences à grande échelle, telles que la campagne d'évaluation des analyseurs syntaxiques EASy (Boullier *et al.* 2005), des expériences d'analyse de corpus de plusieurs millions de phrases, et des expériences d'apprentissage d'informations à partir de corpus de spécialité (classifications sémantiques et construction d'ontologie — projet BIOTIM sur le domaine botanique).

Il est difficile de donner un aperçu quantitatif de la couverture et de la précision du *Lefff*. Toutefois, nous avons développé un « chunker » à règles et reposant sur le *Lefff* qui segmente une phrase en syntagmes non-récursifs (les « constituants » de la campagne EASy), dont les résultats sont satisfaisants (78,5 % de f-mesure) (Sagot 2006 : ch. 12).

Par ailleurs, des travaux sont en cours pour comparer le *Lefff* à différentes autres ressources, et notamment Morphalou (pour le lexique morphologique sous-jacent), mais surtout SynLex, lexique syntaxique cité plus haut, extrait à partir de certaines tables du lexique-grammaire (Gardent *et al.* 2006), et DICOVALENCE, anciennement Proton (van den Eynde et Mertens 2006).

En parallèle à ces enrichissements mutuels et aux travaux d'enrichissement direct tels que ceux décrits dans la section suivante, nous sommes en train de définir une nouvelle architecture pour le *Lefff*, qui prendra mieux en compte la nécessité de construire des interfaces de visualisation, de modification et de correction à la fois souples et simples d'utilisation. Cette nouvelle génération du *Lefff*, la troisième, devrait être disponible dans le courant de l'année 2007.

#### 4. Du lexique-grammaire au *Lefff*

Avant les travaux décrits dans cette section, le *Lefff* était très pauvre quant à la qualité et la quantité de sa couverture des deux grandes classes de constructions verbales non standard que sont les constructions impersonnelles et les expressions verbales figées. La disponibilité d'informations exploitables dans les tables du lexique-grammaire concernant ces constructions nous a amené à envisager leur extraction et leur transformation en vue d'une intégration dans le *Lefff*.

Toutefois, comme indiqué précédemment, il s'agit là d'un travail qui est loin d'être trivial : différents types de travaux sont en effet indispensables pour mener à bien une telle idée :

- Compréhension précise des informations répertoriées dans les tables
- Filtrage des informations linguistiques que l'on souhaite intégrer au *Lefff* (filtrage des colonnes)
- Filtrage des entrées lexicales que l'on souhaite intégrer au *Lefff*
- Développement pour ces constructions d'un modèle compatible avec les principes sur lesquels reposent le *Lefff*
- Extraction effective des informations et des entrées choisies au format *Lefff*, en respectant le modèle retenu
- Si possible, évaluation dans des analyseurs de l'impact d'un tel travail, au moins en termes de taux de couverture.

C'est à cette succession de tâches que nous nous sommes attelés pour les constructions impersonnelles et les expressions verbales figées. Après avoir étudié ces deux cas, nous proposerons une rapide évaluation de l'intérêt de ces travaux.

#### 4.1. Modélisation des impersonnelles dans le *Lefff*

Comme nous l'avons vu dans la section 2.2, une partie du travail avait déjà été effectuée au sujet des impersonnelles lors de la construction de l'outil ILIMP (Danlos 2005). En particulier, la compréhension fine des tables et le filtrage des entrées lexicales étaient déjà réalisés.

Cependant, ILIMP se présente sous forme de réseaux de transition récursifs qui reconnaissent des motifs dans un texte muni d'un étiquetage morphosyntaxique. Ce qui est très différent du point de vue adopté ici, qui est celui de la construction d'un lexique syntaxique, destiné à être utilisé dans un analyseur syntaxique profond. En un sens, notre point de vue est moins complexe, puisque l'on peut faire abstraction (par exemple) de la variabilité du matériau qu'on peut insérer entre une tête verbale et un complément (figé, par exemple), et de la complexité de sa délimitation : c'est en effet aussi le travail de la grammaire, et non seulement du lexique. Mais la contrepartie de cette relative simplification est une plus grande abstraction dans les descriptions. Il ne s'agit plus de décrire des motifs applicables à des séquences de formes étiquetées, mais des cadres de sous-catégorisation et des contraintes syntaxiques.

C'est la raison pour laquelle nous avons construit un modèle des constructions impersonnelles qui n'est pas la transcription directe des graphes d'ILIMP, mais dont le contenu (associations entre lemmes et constructions) en est directement extrait.

La dichotomie entre constructions intrinsèquement impersonnelles et constructions à sujet extraposé (cf. 2.2.2) reste naturellement fondamentale dans notre description. Parmi ces dernières, et en raison de l'importance donnée au cadre de sous-catégorisation fonctionnel, la distinction principale ne se fait pas entre verbes à sujet phrastique et verbes à sujet nominal, mais entre constructions à diathèse active (le sujet extraposé est un sujet profond, il réalise une fonction sujet) et constructions à diathèse passive ou moyenne.

Dans tous les cas, une construction impersonnelle est caractérisée par le fait que la position syntaxique sujet, obligatoire en français dans les propositions finies, est occupée par un pronom impersonnel *il* ou *ça* (mais nous n'avons pas encore traité ce dernier cas). Reprenons successivement les deux classes de constructions impersonnelles identifiées dans la section 2 : les constructions intrinsèquement impersonnelles et les constructions à sujet extraposé.

##### 4.1.1. Les constructions intrinsèquement impersonnelles

Les constructions intrinsèquement impersonnelles sont modélisées par l'absence de fonction sujet dans le cadre de sous-catégorisation. Par conséquent, seule une

construction impersonnelle est possible : c'est ce qui est requis par la macro @impers. Ces constructions sont de trois types<sup>26</sup> :

- Les verbes avec cadre de sous-catégorisation vide (la plupart des verbes météorologiques de la table 31I, comme *il vente*) ;

vente      v    [pred='venter<sub>1</sub>',@impers,@PS3s]

- Les verbes intrinsèquement impersonnels sous-catégorisant une fonction objet, comme les verbes de la table 17 (comme *il faut N/Vinf/queP*) ou la locution *il y a N (Loc N)* ;

faut      v    [pred='falloir<sub>1</sub><Obj:(sn|cla|sinf|scomp)>',@impers,@CompSubj<sup>27</sup>,  
@P3s]

a          v    [pred='y avoir<Obj:sn,Loc:(loc-sn)>',@impers,@pseudo-y,@P3s]<sup>28</sup>

- Les verbes intrinsèquement impersonnels sous-catégorisant une autre fonction (comme *il s'agit de N/Vinf*) ;

agit      v    [pred='s'agir<sub>1</sub><Objde:de-sn|clg|scompl|de-sinf>',@pron,@impers,  
@CompSubj,@P3s]

#### 4.1.2. Les constructions à sujet extraposé

Ces constructions se répartissent entre constructions à prédicat verbal et constructions à prédicat adjectival. Au contraire des constructions intrinsèquement impersonnelles, les cadres de sous-catégorisation incluent une fonction sujet. Il y a donc possibilité d'alternance entre constructions impersonnelles et personnelles.

Comme nous l'avons vu, la macro @impers impose une construction impersonnelle, qui est ici avec sujet extraposé, puisque nous traitons dans cette section des entrées sous-catégorisant une fonction sujet. Les constructions personnelles parallèles, inexistantes pour les constructions intrinsèquement impersonnelles, sont indiquées par la macro @pers. On notera que la duplication au niveau extensionnel des constructions impersonnelles et personnelles n'est pas la conséquence d'une duplication au niveau intensionnel : chaque forme d'une même entrée intensionnelle (du même prédicat) peut être impliquée dans plusieurs constructions, certaines personnelles, d'autres impersonnelles. Comme nous l'avons vu plus haut, c'est déjà le cas pour les verbes personnels, dont le participe passé a deux entrées, l'une active l'autre passive.

Il en est de même pour les adjectifs. On notera qu'il semble que tout adjectif dont la fonction sujet peut avoir une réalisation complétive ou infinitive admette une construction impersonnelle de la forme *il<sub>imp</sub> est Adj Y* (où Y dénote une complétive et/ou une infinitive, suivant les cas). La construction impersonnelle, qui est alors

<sup>26</sup> Pour simplifier la lecture, nous n'avons pas répété l'information cat=v ou cat=adj dans la structure syntaxique des exemples donnés.

<sup>27</sup> Rappelons que cette macro indique que si la fonction objet est réalisée sous la forme d'une complétive, alors celle-ci doit être au subjonctif.

<sup>28</sup> Une autre entrée couvre le cas (familier) *il y a que je suis malade*.

possible, induit une translation du **sinf** en **de-sinf** (*dormir est impossible / il est impossible de dormir*).

On peut regrouper les constructions à sujet extraposé en différentes classes :

- Un certain nombre de verbes admettant une construction impersonnelle en parallèle de la construction personnelle correspondante (toute la table 5, quelques entrées des tables 7 et 8) ;

plaît v [pred='plaire<sub>1</sub><Suj:de-sinf|scompl,Objà:(à-sn|cld)>',@CompSujSubj,  
@impers,@P3s]

plaît v [pred='plaire<sub>1</sub><Suj:sn|cln|sinf|scompl,Objà:(à-sn|cld)>',@pers,@P3s]

découle v [pred='découler<sub>1</sub><Suj:sn|scompl,Objde:de-sn|clg|de-scompl>',  
@DeCompInd,@impers,@PS13s]

découle v [pred='découler<sub>1</sub><Suj:sn|cln|scompl,Objde:de-sn|clg|de-scompl>',  
@DeCompInd,@pers,@PS13s]

- Les verbes non intrinsèquement impersonnels pour lesquels une construction impersonnelle passive ou moyenne existe, soit de façon exclusive, soit à côté d'une construction personnelle (certains verbes, en particulier des tables 6 et 9). L'exemple du participe passé *raconté* permet d'illustrer tous les cas :

- Participe passé actif (*Pierre a raconté un conte à Marie*)

raconté v [pred='raconter<sub>1</sub><Suj:sn,Obj:sn|cla|sinf|scompl,Objà:(à-sn|cld)>',  
@CtrlSuj,@ComplInd,@pers,@Kms]

- Participe passé passif en construction personnelle (*Un conte a été raconté par Pierre à Marie*)

raconté v [pred='raconter<sub>1</sub><Obl:(par-sn),Suj:sn|scompl,Objà:(à-sn|cld)>',  
@passif,@pers,@Kms]

- Participe passé passif en construction impersonnelle (*Il a été raconté un conte à Marie par Pierre*)

raconté v [pred='raconter<sub>1</sub><Obl:(par-sn),Suj:sn|scompl,Objà:(à-sn|cld)>',  
@passif,@impers,@Kms]

- Participe passé moyen en construction personnelle (*Un (tel) conte (ne) s'est (pas) raconté à (quelqu'un comme) Marie (depuis longtemps)*)

raconté v [pred='raconter<sub>1</sub><Suj:sn|scompl,Objà:(à-sn|cld)>',@pron,@impers,  
@Kms] # se moyen impersonnel

- Participe passé moyen en construction impersonnelle (*Il (ne) s'est (pas) raconté un (tel) conte à (quelqu'un comme) Marie (depuis longtemps)*)

raconté v [pred='raconter<sub>1</sub><Suj:sn|scompl,Objà:(à-sn|cld)>',@pron,@pers,  
@Kms] # se moyen personnel

- Un certain nombre d'adjectifs admettant une construction impersonnelle de type *ilimp est Adj Y* (*il est facile de faire cela*). On notera qu'un adjectif sous-



catégorise toujours une fonction sujet, et parfois d'autres fonctions également<sup>29</sup> ;

envisageable adj [pred='envisageable<sub>1</sub><Suj:(de-sinf|scompl)>',@impers,@s]

envisageable adj [pred='envisageable<sub>1</sub><Suj:(sn|sinf|scompl)>',@pers,@s]

- Un certain nombre d'expressions de type être Prep X admettant une construction impersonnelle de type ilimp est Prep X Y (*il est de règle de porter un chapeau*). Elles sont similaires à des adjectifs, et catégorisées comme telles.

d'actualité adj [pred='d'actualité<sub>1</sub><Suj:(de-sinf|scompl)>',@impers,@s]

d'actualité adj [pred='d'actualité<sub>1</sub><Suj:(sn|sinf|scompl)>',@pers,@s]

à l'actif adj [pred='à l'actif<sub>1</sub><Suj:(de-sinf|scompl),Objde:de-sn>',@impers,@s]

à l'actif adj [pred='à l'actif<sub>1</sub><Suj:(sn|sinf|scompl),Objde:de-sn>',@pers,@s]<sup>30</sup>

#### 4.1.3. Bilan

Nous avons donc extrait des différents graphes pertinents qui constituent ILIMP des listes de verbes et d'adjectifs associés à chacun de ces cas. Puis nous avons créé de nouvelles classes syntaxiques pour les constructions impersonnelles, ou modifié certaines classes existantes, afin d'ajouter ou de modifier les entrées du *Lefff* extensionnel d'une façon cohérente à la fois avec l'analyse ci-dessus et les principes de représentation présentés en section 3.

#### 4.2. Modélisation des expressions verbales figées dans le *Lefff*

Comme nous l'avons vu, la prise en compte des constructions impersonnelles dans le *Lefff* à partir des tables du lexique grammairal (via ILIMP) a conduit à un résultat raisonnablement exhaustif, selon une modélisation linguistiquement satisfaisante. Pour les expressions verbales figées, la situation est plus complexe, en raison de plusieurs phénomènes de natures variées. Tous contribuent à rendre difficile la prise en compte dans le *Lefff* de ces expressions :

- Seule environ une moitié des tables des expressions verbales figées sont actuellement disponibles sous licence LGPL-LR
- Comme indiqué précédemment, il existe un continuum entre expressions verbales libres et expressions verbales complètement figées
- Les phénomènes de figement induisent des contraintes complexes, par exemple en termes d'ordre des mots, de modifiabilité des compléments constants, de

<sup>29</sup> On notera que l'information sur le mode de la complétive n'est actuellement pas disponible. Il en est de même ci-dessous pour les expressions en être Prep X.

<sup>30</sup> Anticipant ainsi certains problèmes liés au figement, notons que ces entrées ne couvrent pas le cas à son actif. Il est donc nécessaire d'introduire, indépendamment, un lemme à son actif dont les formes fléchies (à mon actif, à ton actif,...) ne sous-catégorisent qu'un sujet.

coréférence entre compléments, de variabilité paradigmaticque limitée mais non nécessairement nulle, etc.

- Tout figement implique une séparation entre structure interne et structure externe de l'expression, séparation dont l'intrinsèque complexité est augmentée par sa perméabilité à certains phénomènes.

En réalité, il n'existe pas, à l'heure actuelle, de modèle entièrement satisfaisant qui permettrait de traiter l'ensemble des phénomènes regroupés sous le terme d'expressions figées. Même en imaginant un modèle très simple, on peut être amené à proposer, suivant les cas :

- une approximation de type *constituants figés*, où un figement est caractérisé par la non-variabilité d'une portion connexe de l'arbre de constituant d'une phrase, ainsi éventuellement que de contraintes complémentaires,
- une approximation de type *dépendances figées*, où un figement est caractérisé par la non-variabilité d'une portion connexe de l'arbre de dépendances d'une phrase, ainsi éventuellement que de contraintes complémentaires.

Le traitement que nous avons réservé aux classes d'expressions verbales figées est de type *constituants figés* : nous modélisons une telle expression en attribuant une structure unique non-réursive (un seul « pred ») à la séquence du ou des constituants figés, considérée comme strictement invariable, et qui indique explicitement quelle est la nature de la tête verbale requise. Ceci permet l'insertion, entre la tête verbale et le(s) complément(s) figé(s), de différents matériaux (modifieurs, compléments sous-catégorisés par l'expression,...).

Par exemple, *prendre en compte (qqch)* sera modélisé principalement par l'entrée suivante (comme indiqué au tableau 1, « cf » est le terminal attribué aux « compléments figés » dont le placement par rapport à d'éventuels autres compléments est libre) :

en compte      cf    [pred='prendre en compte<sub>1</sub><Suj:sn|cln,Obj:(sn|cla)>',  
                             synt\_head=prendre]

Une telle entrée suit un modèle similaire à l'entrée d'un nom prédicatif, qui peut se construire avec un verbe support (Danlos 1992). Le modèle utilisé pour les expressions verbales figées est en effet très proche, puisque dans les deux cas un complément non-libre impose une tête verbale fixe (modulo, dans certains cas, des variantes aspectuelles).

Les limites de ce modèle sont liées à l'absence de structure interne attribuée au(x) complément(s) figé(s) : ceci a pour effet

- d'interdire le défigement,
- d'empêcher une modélisation pleinement satisfaisante des réalisations infixes de certains compléments (comme vu précédemment, à l'actif de *qqn* / à son actif),

- de rendre très difficile la modélisation de l'insertion de matériau entre deux compléments figés corrélés l'un à l'autre,
- de réserver pour un traitement spécifique le découpage en unités inférieures (chunks, « compléments » lorsque plusieurs sont figés).

Malgré ces limites, nous avons pris en compte plusieurs tables d'expressions verbales figées dans le *Lefff*, ce qui permet une analyse raisonnablement satisfaisante des phrases qui les mettent en œuvre.

#### 4.3. Exemples d'utilisation dans un analyseur LFG

En parallèle à la modélisation dans le *Lefff* des constructions impersonnelles et de certaines expressions verbales figées, les analyseurs syntaxiques reposant sur le *Lefff* ont fait évoluer leurs grammaires, afin de pouvoir bénéficier de ces informations nouvelles et analyser correctement les phrases où elles sont mises en œuvre.

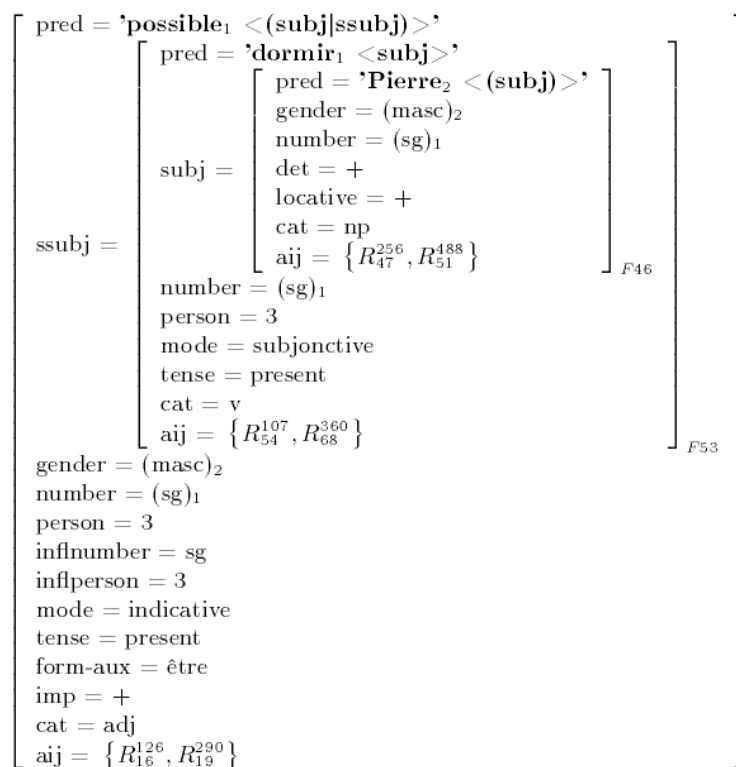


Figure 1. Structure fonctionnelle LFG produite pour la phrase (10a) par un analyseur reposant sur le *Lefff* 2.2.2.

Ce n'est pas l'objet de cet article que de présenter ces grammaires et ces analyseurs en détail, pas plus que de proposer une évaluation qualitative à grande échelle sur corpus de l'amélioration obtenue grâce à la prise en compte des phénomènes que nous avons étudié. Toutefois, à titre d'illustration, nous proposons deux exemples d'analyses de phrases simples, l'un avec une construction impersonnelle (figure 1, phrase (10a)), l'autre avec une expression verbale figée (figure 2, phrase (10b)). Les résultats montrés sont produits par l'analyseur que construit le système SXLFG à partir de la

grammaire du français qui lui est associée. Il s'agit donc d'analyses LFG, dont nous ne montrons que les structures fonctionnelles. On remarquera, comme dit plus haut, que le format du *Lefff* a été adapté au formalisme LFG avant son intégration dans l'analyseur.

- (10)a. Il est possible que Pierre dorme.  
b. Pierre prend en compte ce problème.

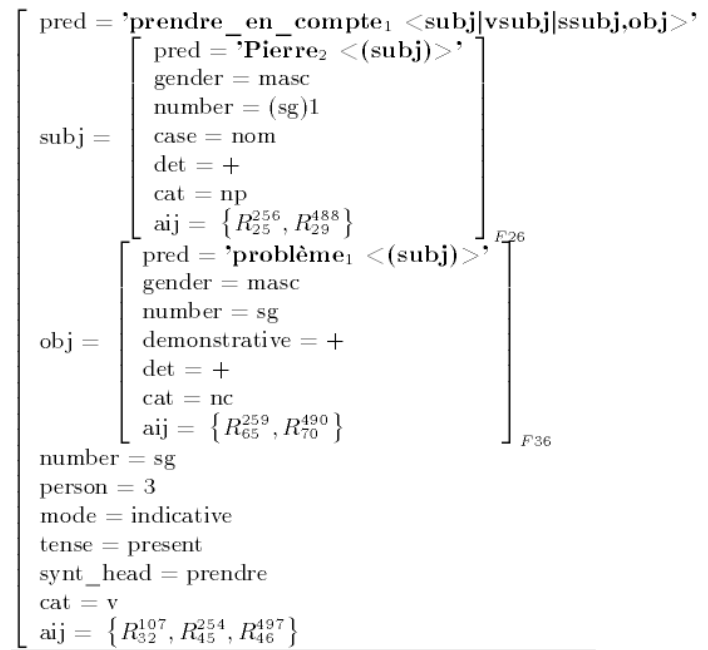


Figure 2. Structure fonctionnelle LFG produite pour la phrase (10b) par un analyseur reposant sur le *Lefff* 2.2.2.

## Conclusion

Le *Lefff*, lexique syntaxique du français à large couverture, a désormais des fondements linguistiques et formels qui permettent son utilisation dans des analyseurs syntaxiques profonds à grande échelle. Ces fondements nous ont permis d'exploiter la source d'informations linguistique que sont les tables du lexique-grammaire, pour modéliser deux familles de phénomènes syntaxiques non standard : les constructions impersonnelles et certaines expressions verbales figées. Le travail sur les constructions impersonnelles semble satisfaisant, mais celui sur le figement reste préliminaire, il nous faut le poursuivre. De plus, la mise à disposition du lexique DICOVALENCE, successeur du lexique PROTON, nous permet de disposer d'une autre source d'informations linguistiques, ce qui sera particulièrement utile, entre autres, pour modéliser de façon satisfaisante l'ensemble des constructions pronominales, aujourd'hui à l'état d'ébauche dans le *Lefff*.

## Références

- BOONS J.-P., GUILLET A., LECLERE C. (1976a), La structure des phrases simples en français, Constructions intransitives, Droz, Genève.
- BOONS J.-P., GUILLET A., LECLERE C. (1976b), *La structure des phrases simples en français, Classes de constructions transitives*, Rapport de recherches, LADL, CNRS, Univ. Paris 7.
- BOULLIER P. et SAGOT B. (2005), « Analyse syntaxique profonde à grande échelle: SxLFG », in *Traitement Automatique des Langues*, n° 46/2.
- CLEMENT L., SAGOT B. et LANG B. (2004), « Morphology based automatic acquisition of large-coverage lexica », in *Proceedings of LREC 2004*, Lisbonne, Portugal.
- DANLOS L. (1980), Représentation d'informations linguistiques: les constructions N être Prép X, Thèse de troisième cycle, Université Paris 7.
- DANLOS L. (1992), « Support Verb Constructions: linguistic properties, representation, translation », in *Journal of French Linguistic Studies*, n° 2/1, Cambridge University Press, Cambridge.
- DANLOS L. (2006), « ILIMP : Outil pour repérer les occurrences du pronom impersonnel *il* », in *Actes de TALN 2006*, Louvain, Belgique.
- VAN DEN EYNDE K. et BLANCHE-BENVENISTE, C. (1978), « Syntaxe et mécanismes descriptifs : présentation de l'approche pronominale », in *Cahiers de Lexicologie* n°32 : 3-27.
- VAN DEN EYNDE K. et MERTENS P. (2006), Le dictionnaire de valence DICOVALENCE : manuel d'utilisation, à paraître.
- GARDENT C., GUILLAUME B., PERRIER G. et FALK I. (2006), « Extraction d'information de sous-catégorisation à partir des tables du LADL », in *Actes de TALN 2006*, Louvain, Belgique.
- GROSS M. (1982), « Une classification des phrases “figées” du français », in *Revue Québécoise de linguistique*, 11/2 : 151-185.
- GROSS M. (1993), « Les phrases figées en français », in *L'information grammaticale*, n° 59, Paris : 36-41.
- GUILLET A. et LECLERE C. (1992), La structure des phrases simples en français, Constructions transitives locatives, Droz, Genève.
- LAMIROY B., LECLERE C., KLEIN J. R., LABELLE J. (2003), « Expressions verbales figées et variation en français: le projet BFQS », in *Cahiers de lexicologie*, n° 83/2 : 153-172.
- LAMIROY B. et KLEIN J. R. (2006), « Le problème central du figement est le semi-figement », in *Linx*, à paraître.
- SAGOT B. (2005), « Automatic acquisition of a Slovak Lexicon from a Raw Corpus », in *Proceedings of TSD 2005*, Karlovy Vary, République Tchèque (Lecture Notes in Artificial Intelligence 3658, © Springer-Verlag).
- SAGOT B., CLEMENT L., DE LA CLERGERIE É. et BOULLIER P. (2006), « The Lefff 2 syntactic lexicon for French: architecture, acquisition, use », in *Actes de LREC 2006*, Gênes, Italie.
- SAGOT B. et DE LA CLERGERIE É. (2006), « Error mining in parsing results », in *Proceedings of ACL-CoLing 2006*, Sydney, Australie.
- SAGOT B. (2006), Analyse automatique du français : lexiques, formalismes, analyseurs, Thèse de doctorat, Université Paris 7.
- VERONIS J. (1998), Multext-Lexicons, A set of Electronic Lexicons for European Languages.
- THOMASSET F., DE LA CLERGERIE É. (2005), « Comment obtenir plus des méta-grammaires », in *Actes de TALN 2005*, Dourdan, Belgique.